AD________________

Award Number: DAMD17-96-1-6226

TITLE: Computer-Aided Diagnosis of Breast Cancer: A Multi-Center Demonstrator

PRINCIPAL INVESTIGATOR: Carey E. Floyd, Jr., Ph.D.

CONTRACTING ORGANIZATION: Duke University Medical Center
Durham, North Carolina 27710

REPORT DATE: October 1999

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
October 1999

3. REPORT TYPE AND DATES COVERED
Annual (1 Oct 98 - 30 Sep 99)

4. TITLE AND SUBTITLE Computer-Aided Diagnosis of Breast Cancer: A Multi-Center Demonstrator

5. FUNDING NUMBERS
DAMD17-96-1-6226

6. AUTHOR(S)
Carey E. Floyd, Jr., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Duke University Medical Center
Durham, North Carolina 27710

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The long range goal of this project is to improve the accuracy and consistency of breast cancer diagnosis by developing a Computer Aided Diagnosis (CAD) system for early prediction of breast cancer from the patient's mammographic findings and medical history.

In the fifth year of this project, we have acquired 400 new cases which bring our total case database to over 2500. A user interface (developed last year) was implemented for efficient data entry with error checking. These developments continued the first specific aim of the grant: develop an ANN to predict biopsy outcome from mammographic and history findings. In this year, we continued to focus on the second specific aim: evaluate the improvement in radiologists' diagnostic performance when the computer diagnostic aid is provided. This implementation of an accurate CAD system will improve sensitivity, specificity, and consistency of breast cancer diagnosis and will provide a significant improvement in a long-term outcome for breast cancer patients.

14. SUBJECT TERMS Breast Cancer

15. NUMBER OF PAGES
20

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

Carey E [illegible] 25 Feb 00

PI - Signature Date

Table of Contents

Report of the progress on Grant DAMD17-96-1-6226

for the period October 1998 to October 1999.

Introduction

Mammography is the most sensitive procedure for detecting breast cancer. Unfortunately, as currently practiced, the positive predictive value (PPV) is low. Between 70% and 90% of women who undergo biopsy for mammographically suspicious nonpalpable lesions have no malignancy[1] Between 0.5 - 2.0% of all mammographic exams result in biopsy. Each year this amounts to several hundreds of thousands of biopsies performed on benign lesions. Women who undergoing biopsy for a benign finding are unnecessarily subjected to the discomfort, expense, potential complications, change in cosmetic appearance, and anxiety that can accompany breast biopsy[1-4]. The cost of these procedures is between $3000 and $5000 per biopsy and is significant in the present political and economic effort to reduce expenditures. In clinical practice, mammography reporting systems are typically implemented as a data entry form into a relational data base. The system that we describe in this report can be easily integrated into the mammographers' work-flow since it is also based on a relational database structure. The clinician interprets the mammogram, records the findings using a standard reporting lexicon (BI-RADS^{TM}), and enters these findings into the database. All of this is currently the standard procedure. The database is searched for similar cases and the fraction of those similar cases that were malignant is returned. In practice, a threshold is applied to the fraction and if the fraction is above the threshold, the computer aid would recommend biopsy. The woman's health care team can then include this recommendation in the medical decision for biopsy. The long term hope is that this computer aided approach may significantly improve the delivery of health care to these women. A long-term goal of this project has been to gather mammographic data from multiple sites in order to verify and whether the artificial neural network computer aid to the diagnosis of breast cancer can be translated between locations. While the system has proven to be robust and could in principle be trained for every application location, much facility could be gained if we could demonstrate that a single System could be developed and deployed nationally. This deployment would facilitate transferring the expertise currently present in only a few tertiary care centers to the public at large and to smaller and more rural settings.

Progress

Progress in this funding period: October 1998 to October 1999 is demonstrated through the 8 publications supported in part by this grant.

The publications included four peer-reviewed journal articles (2 published, 2 in press), three manuscripts in conference proceedings, and two conference abstracts for presentations at the RSNA meeting.

Peer-reviewed Journal Manuscripts

Baker JA, Frederick ED, Lo JY, Kornguth PJ, and Floyd CE Jr. Incorporation of an Artificial Neural Network into Clinical Mammography to Reduce Benign Breast Biopsies. *AJR Supplement* 170:84, 1998.

Lo JY, Baker JA, Kornguth PJ, Floyd CE Jr. Effect of Patient History Data on the Prediction of Breast Cancer from Mammographic Findings with Artificial Neural Networks. *Acad Radiol*, 6;10-15; 1999.

Floyd CE Jr, Lo JY, Tourassi GD. Breast Biopsy: Case-Based Reasoning Computer-Aided Using Mammography Findings for the Decision to Biopsy. In Press to *American Journal of Roentgenology* (AJR) [5/99].

Gavrielides MA, Lo J, Vargas-Voracek R, Floyd CE Jr. Segmentation of suspicious clustered microcalcification in mammograms. *Medical Physics.*

Conference Proceedings Manuscripts

Vargas-Voracek R, Floyd CE Jr. Computer-Aided Diagnosis for Early Detection of Breast Cancer from Mammograms. Susan G. Komen Breast Cancer Foundation "Reaching for the Cure" National Grant Conference. (1998).

Vargas-Voracek R, Floyd CE Jr. Hierarchical Markov-Random Field Texture Modeling for Mammographic Structure Segmentation Using Multiple Spatial and Intensity Image Resolutions. Presented at the 1999 Medical Imaging Symposium. International Society for Optical Engineering (SPIE). February 20-26, 1999.

Tourassi GD, Floyd CE Jr, Lo JY. A Constraint Satisfaction Neural Network for Medical Diagnosis. Presented at the 1999 International Conference on Neural Networks (ICNN), Washington, DC.

Conference Abstracts

Vargas-Voracek R, Floyd CE Jr. Markov-Random Field Texture Model for Automatic Breast Parenchyma Characterization. Presented at the 84th Radiological Society of North America (RSNA) Scientific Assembly and Annual Meeting. November 29-December 4, 1998.

Lo JY, Kornguth PJ, Floyd CE Jr. Multi-Institution Evaluation of BIRADS Breast Cancer Prediction Model. *Radiology* 209(P):271, 1998.

Methods

To assess how the proposed systems might perform in different health care selivery settings, we have acquired a mammographic feature set along with biopsy outcomes from three different institutions. With a another two institutions data arriving in the next two months, currently, we have over 2500 cases for testing and training and evaluating the artificial intelligence computer aids. With the new data that is about to arrive, this total will rise to over 5000 cases. During the research performed under this application we have discovered several important things about database research. These cases are difficult to obtain. While there are a number of investigators who would be able to provide the mammographic data alone, the need for patient demographic data dramatically increases the amount of research effort required to obtain the data that we need. Several of our original collaborators found that they were unable to support the research effort required with the funds that were provided to us for this task. Our initial estimates of the financial cost of providing and acquiring these cases was an underestimate. This is due in part to the rapid evolution of economic restructuring in major research medical centers over the last five years. While the overall result of this restructuring on the medical health care economic situation has been positive, the impact on research has been very negative. The very simple explanation is that hospitals are no longer able to provide a level of infrastructure supporting previously afforded to research activities. The impact of this on this research project is that the acquisition of cases, so critical to this project, is more expensive than anticipated.

At our own institution , Duke University, we have established an accurate and efficient procedure for obtaining the mamographic data, the pathological data, and the demographic data. It is unfortunate that the integrated medical radiological information system that was scheduled to go on line within the first year has yet to be realized. Nonetheless, and through diligent application of old-fashioned data acquisition using paper forms and hand verification, we have acquired over a thousand cases which have been extensively verified. These 1027 cases, combined with the 996 cases from the University of Pennsylvania, form our current data set. A preliminary evaluation of the similarities and differences between the data sets acquired at the three medical institutions is presented here.

Over the last year we have performed several comparisons of a neural network and other classification systems on the three data sets. Software has been developed to facilitate the rapid organization and comparison of multiple data sets and to facilitate the combination of these data sets into training, testing, and evaluation sets. In an earlier progress report, we demonstrated that the distributions of mammographic findings do not adhere to a normal distribution pattern. Particularly, this is true given the relatively small number of cases in any one finding. Accepting this reality, there are few statistical tests that are appropriate to apply when trying to describe the similarities and differences between the distributions of findings over the three institutions. One technique that is rigorous and at the same time intuitively appealing, is that of case matching. With this technique we set definitions of similarity and then search for cases which are similar between the two data sets given these definitions. The definitions may be strict or maybe lax and the failure or success of the similarity matching under these different criteria can form the basis for describing the similarity of the two data sets. This is in fact an implementation of the artificial intelligence classification technique known as cases based reasoning.

We implemented a case based reasoning formalism using the Microsoft access database language. In fact, after implementing the system as a technique for comparing the databases, we found that it did in itself make a very good classifier . It is in this form that we have implemented the case based reasoning and applied it to the task of determining similarity or difference between the study databases . Below we present the results of the the evaluation of these data sets using the case based reasoning system under a reasonably lax matching criteria.

The case based reasoning algorithm is very simple and intuitive. Case based reasoning is a computer implementation of the question "of all the cases in one data set, how many match a particular selected case from another data set." To investigate this question, the two data sets are structured as tables in a database and sequel query language is employed to perform the matching and scoring. Matching rules are implemented as numerical and logical conditions for the query calls. The results set from this query is a list of all cases in the reference database that matched the single case selected from the test data set. A malignancy ratio is formed as the ratio of all cases in the match list which were malignant at biopsy/the total number of cases that matched. This process is repeated for each case in the test data set. The malignancy ratio is taken as a decision variable and the R O C performance is evaluated. An evaluation of the similarity of the two data sets may be obtained by switching the roles of the data sets in this process. The data set which was initially used as the reference data set is now used as the test data set while the data set which was originally used as the test set is now used as the reference. Comparison of the two R O C results forms a functionally useful test for similarity. The goal of this evaluation was into determined if, as used in the computer aided prediction models, the two data sets were equivalent.

Results

The ROC plot shown in Fig. 1 , demonstrates the performance for predicting the outcome of biopsy when the Duke data set is used as the testing set. Results for two different reference databases are plotted. The solid line shows the results when the Duke data set is used as the reference database while the dashed line shows the performance when the Penn data set is used as the reference database. While there is almost no difference in the area under the two curves, the model that used the Duke data set as the reference database has higher performance in the region of high sensitivity which is where the system would operate in a clinical application.

The ROC plot shown in Fig. 2 , demonstrates the performance for predicting the outcome of biopsy when the Penn data set is used as the testing set. Results for the two different reference databases again are plotted. The solid line shows the results when the Duke data set is used as the reference database while the dashed line shows the performance when the Penn data set is used as the reference database. The model using the Duke reference database shows higher performance for sensitivities between 80 and 90%, but each reference database provides equally good performance for sensitivities from 90% to 100%.

Both the Duke and Penn data have been explored using an ANN model. For comparison, the performances of the ANN and both CBR models are shown in Fig. 3 when predicting the outcomes for the Duke data set. The solid line shows the results for CBR when the Duke data set is used as the reference database while the dashed line shows the performance when the Penn data set is used as the reference database. The dotted line shows the performance of the ANN for comparison. Of the three predictive models, CBR using the Duke reference database shows the highest performance for sensitivities between 98 and 100%, although the difference is very small and is not statistically significant. For sensitivities between 98% and 75%, the ANN provides higher performance than either of the CBR models.

The matrix in Table 1 shows the predictive performance as an ROC area for all combinations of the Duke, Penn, and the combined (noted as "Both") data sets. Each testing data set is specified by a column with the name listed in the first row. Each reference database is specified by a row with the name listed in the first column. The corresponding ROC area is located at the intersection. There is essentially no difference in the performance for predicting the outcomes in any of the three data sets regardless of which is used as the reference.

Table 2. presents a comparison of several measures of the predictive performance for both the Duke and Penn testing data sets using both the CBR and the ANN models. Here, the ROC area is compared with the specificity at 98% sensitivity. In addition, the performance is presented for this threshold setting that produces 98% sensitivity as the number of benign biopsies that could have been spared along with the number of malignancies that would have been missed.

Conclusion

From the study just described we can conclude that the data sets from the University of Pennsylvania and from Duke University are equivalent in terms of similarity in the distributions of BIRADS findings and their relationship to the likelihood of malignancy. While not quantitatively analyzed, it seems intuitively obvious that there could be differences between the patient populations from these data sets. With the set from U of Penn (Philadelphia) representing an urban population and that from Duke (Durham) representing a more rural population. That these differences were not seen in the experiments suggests that this aspect of patient population may not be a factor. Particularly, the similarity between U of Penn and Duke would suggest that the predictive model described here is relatively insensitive to the differences in these patient populations. These results are supportive of the conclusion that a separate predictive model for each intended local may not be required.

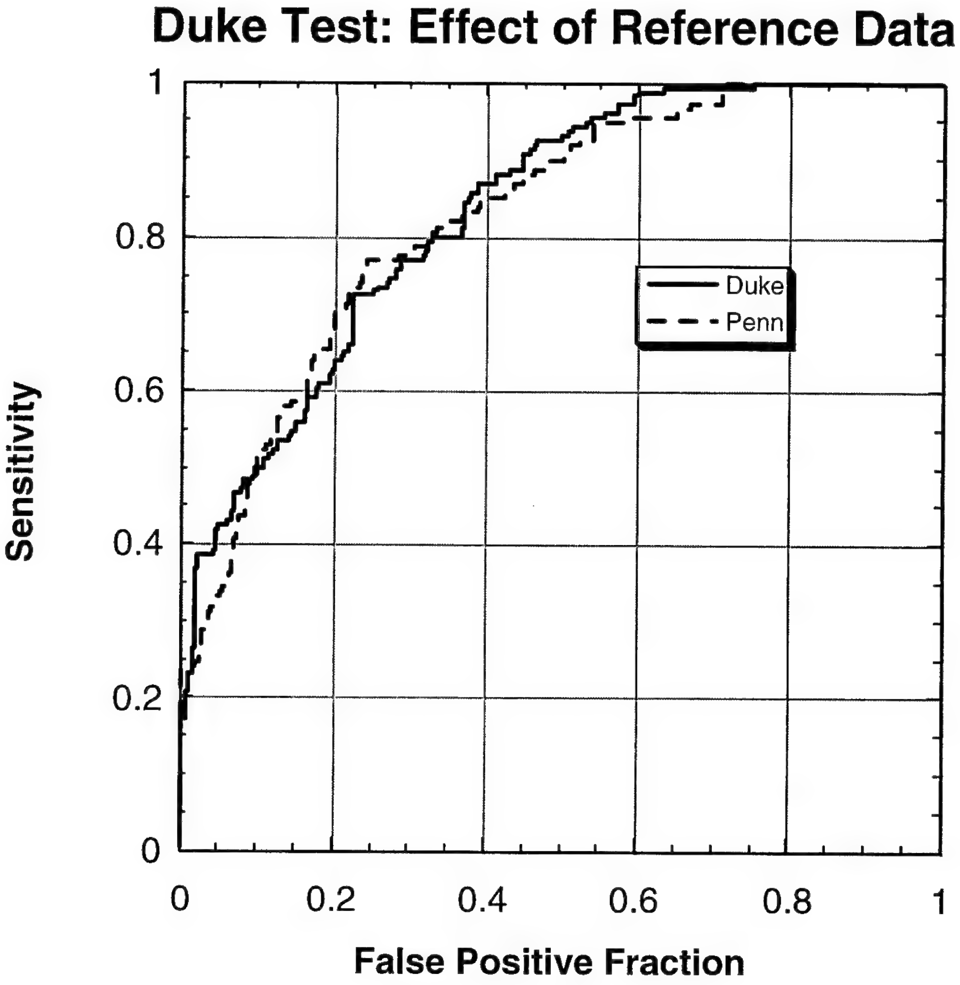

Figure 1.The ROC plot of the predictive model performance for predicting the outcome of biopsy when the Duke data set is used as the testing set. The solid line show the results when the Duke data set is used as the reference database while the dashed line shows the performance when the Penn data set is used as the reference database. The Duke reference database shows higher performance in the region of high sensitivity.

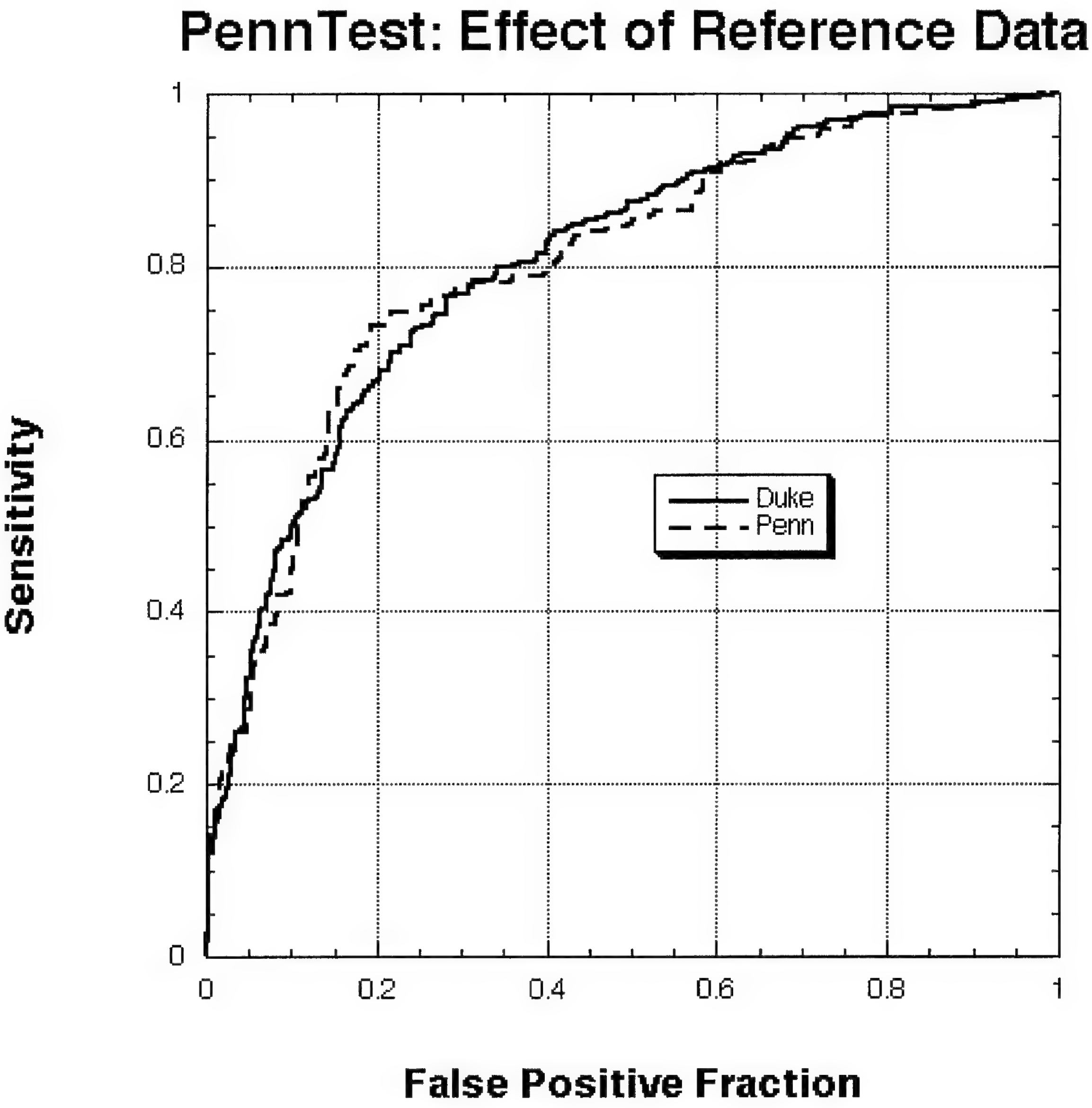

Figure 2.The ROC plot of the performance for predicting the outcome of biopsy when the Penn data set is used as the testing set. The solid line show the results when the Duke data set is used as the reference database while the dashed line shows the performance when the Penn data set is used as the reference database. The Duke reference database shows higher performance for sensitivities between 80 and 90%, but each reference database provides equally good performance for sensitivities from 90% to 100%.

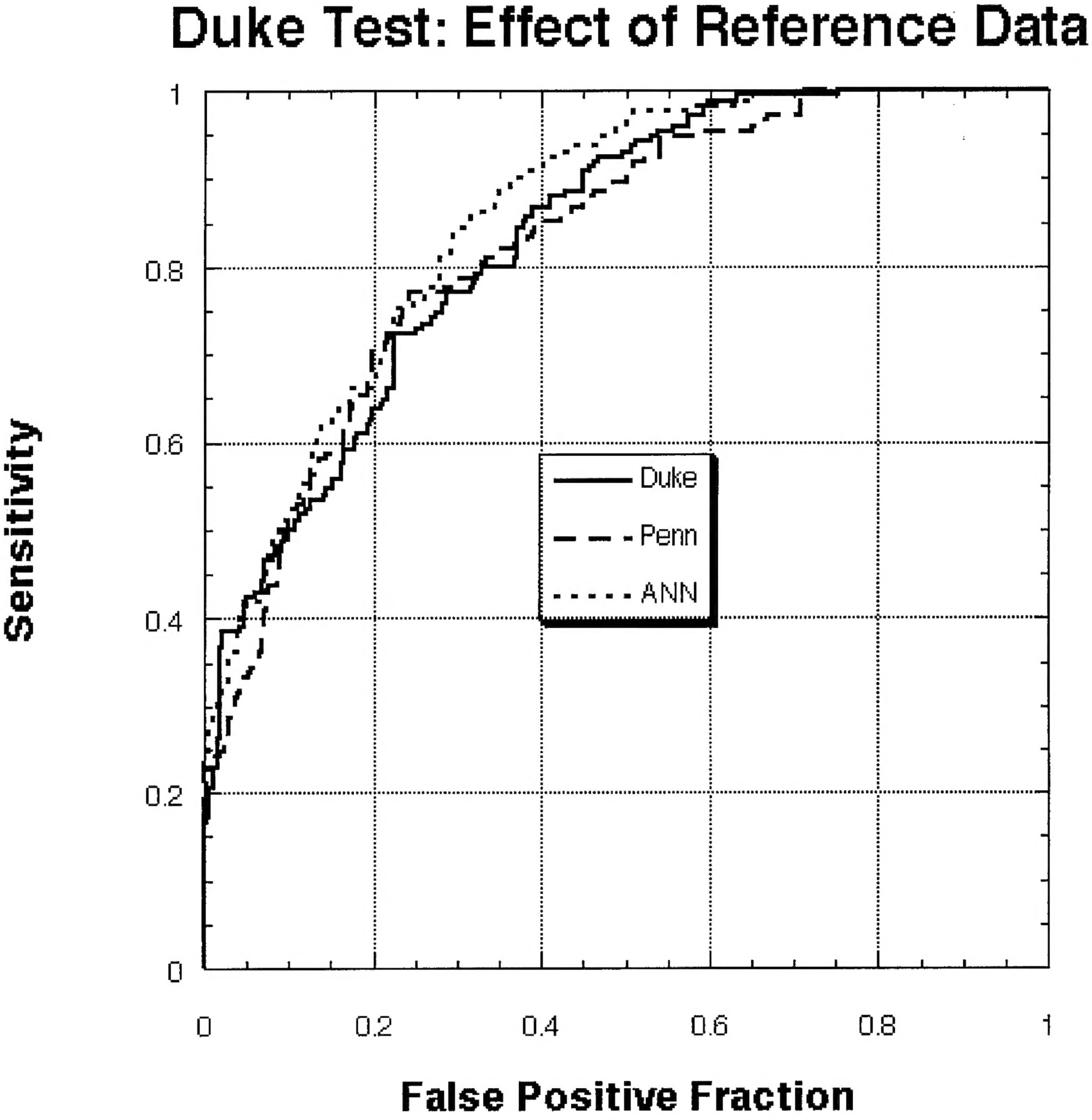

Figure 3.The ROC plot of the performance for predicting the outcome of biopsy when the Duke data set is used as the testing set. The solid line show the results for case-based reasoning when the Duke data set is used as the reference database while the dashed line shows the performance when the Penn data set is used as the reference database. The dotted line shows the performance of an artificial neural network for comparison. Of the three predictive models, case-based reasoning model using the Duke reference database shows the highest performance for sensitivities between 98 and 100%, although the difference is very small and is not statistically significant. For sensitivities between 98% and 75%, the artificial neural network provides higher performance than either of the case-based reasoning models.

		Testing Data set		
		Duke	Penn	Both
Reference Database	Duke	0.83	0.81	0.81
	Penn	0.83	0.80	0.81
	Both	0.83	0.80	0.82

Table 1. The matrix in Table 1 shows the predictive performance as an ROC area for all combinations of the Duke, Penn, and the combined (noted as "Both") data sets. Each testing data set is specified by a column with the name listed in the first row. Each reference database is specified by a row with the name listed in the first column. The corresponding ROC area is located at the intersection. There is essentially no difference in the performance for predicting the outcomes in any of the three data sets regardless of which is used as the reference.

Comparison of performance of CBR with ANN

Testing Data - Model	ROC area	Specificity at 98% Sensitivity	Benign: Spared/ Total	Malignant: Missed/Total
Duke-CBR	0.83	0.41	134/326	3/174
Penn-CBR	0.80	0.17	103/603	7/394
Duke-Ann	0.86	0.42	136/326	3/174
Penn -Ann	0.82	0.15	90/603	7/394

Table 2. A comparison of several measures of the predictive performance for both the Duke and Penn testing data sets using both the CBR and the ANN models. (CBR=case-based reasoning, ANN=artificial neural network)

References

1. Kopans DB. The positive predictive value of mammography. AJR. American Journal of Roentgenology 1992; 158: 521-526
2. Dixon JM and John TG. Morbidity after breast biopsy for benign disease in a screened population. Lancet 1992; 1: 128
3. Helvie MA, Ikeda DM, and Adler DD. Localization and needle aspiration of breast lesions: complications in 370 cases. AJR. American Journal of Roentgenology 1991; 157: 711-714
4. Schwartz GF, Carter DL, Conant EF, Gannon FH, Finkel GC, and Feig SA. Mammographically detected breast cancer: nonpalpable is not a synonym for inconsequential. Cancer 1994; 73: 1660-1665